Vision in Action: Learning Active Perception from Human Demonstrations

Haoyu Xiong **Xiaomeng Xu** **Jimmy Wu** **Yifan Hou** **Jeannette Bohg** **Shuran Song**
Stanford University
https://vision-in-action.github.io

Abstract: We present Vision in Action (ViA), an active perception system for bimanual robot manipulation. ViA learns task-relevant active perceptual strategies (e.g., searching, tracking, and focusing) directly from human demonstrations. On the hardware side, ViA employs a simple yet effective 6-DoF robotic neck to enable flexible, human-like head movements. To capture human active perception strategies, we design a VR-based teleoperation interface that creates a shared observation space between the robot and the human operator. To mitigate VR motion sickness caused by latency in the robot's physical movements, the interface uses an intermediate 3D scene representation, enabling real-time view rendering on the operator side while asynchronously updating the scene with the robot's latest observations. Together, these design elements enable the learning of robust visuomotor policies for three complex, multi-stage bimanual manipulation tasks involving visual occlusions, significantly outperforming baseline systems.

Keywords: Active Perception, Bimanual Manipulation, Imitation Learning, Teleoperation Systems

1 Introduction

Perception is inherently active [1]. Consider the task of retrieving a banana from a bag (Fig. 1): one must first scan the environment to locate the bag, then peek inside to identify the banana, and finally focus on the object to determine an appropriate grasp. These deliberate viewpoint changes serve to **increase** visual coverage during the search, **reduce** occlusions caused by obstacles (*e.g.*, the bag), and **focus** attention on action-critical regions (*e.g.*, for grasp finding).

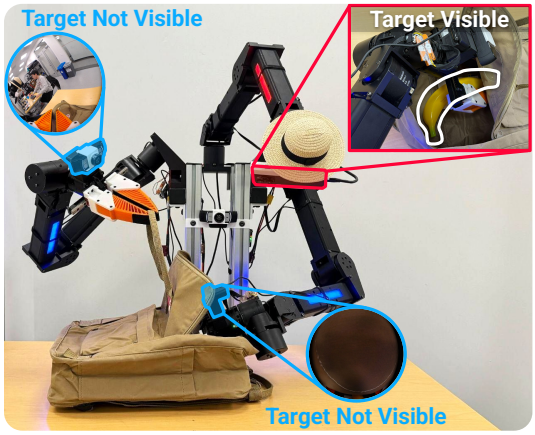

Figure 1: **Vision in Action (ViA)** uses an active head camera to search for the target object (yellow banana) inside the bag. The wrist cameras are ineffective in this visually occluded scenario, as they are constrained by the arm motions.

Yet, most robotic imitation learning systems [2, 3, 4, 5, 6, 7, 8, 9] do not incorporate active perception. These systems typically rely on wrist cameras [6, 7, 8] or fixed third-person cameras [10]. Since wrist cameras move with the arm, their viewpoints are constrained by manipulation requirements rather than guided by perceptual objectives. This limitation becomes especially problematic in scenarios involving visual occlusion, where wrist cameras are often blocked by the environment and fail to capture task-relevant information necessary for accurate action inference. Furthermore, during data collection, humans naturally shift their gaze to guide attention. However, the robot usually perceives the scene from fixed or mismatched viewpoints. As a result, these systems fail to capture rich human perceptual behaviors such as searching, tracking, and focusing. This fundamental **observation mismatch**—

For any questions, please contact: haoyux.me@gmail.com

between *what the human sees* and *what the robot learns from*—introduces a critical gap that ultimately hinders the learning of effective policies. Despite its importance, active perception is often neglected in today's robotic systems due to the significant system-level challenges it introduces, including:

- *Flexible hardware for human-like gaze control.* While humans effortlessly coordinate eye, neck, and torso movements to direct their gaze in a variety of ways, replicating this capability in robots is difficult. Most robot systems today rely on fixed or constrained cameras (*e.g.*, 2-DoF necks [11, 12, 13, 14]), which limit the ability to adjust viewpoints flexibly.
- *Synchronized camera-gaze movements.* Virtual reality provides a powerful interface for teleoperating robots and capturing human active perception [14, 15]. However, designing an interface that synchronizes human gaze and movement of the robot camera requires precise mirroring of human motions and real-time streaming of visual feedback. Achieving this demands fast motor control and low-latency data streaming, both of which remain challenging with today's hardware.
- *Scalable active perception strategies.* Human gaze is driven by top-down and bottom-up attention [16, 17, 18, 19]. Prior efforts to replicate human gaze behavior in robots typically relied on hand-crafted heuristics [20, 21, 22, 23], but such strategies are difficult to generalize across diverse tasks. A more scalable approach should allow the robot to learn active perception strategies that maximize task-relevant information gain, without requiring task-specific assumptions.

In this paper, we introduce **Vision in Action (ViA)**, a bimanual manipulation system that learns active perception strategies directly from human demonstrations. Our system addresses the above challenges using the following design choices:

- *Flexible robot neck using an off-the-shelf 6-DoF arm.* Instead of replicating the intricate biomechanics of the human neck and torso through a complex design, we use an off-the-shelf 6-DoF robot arm as the robot's neck. This simple yet effective approach enables human-like head movements that approximate the full range of motion produced by coordinated upper-body motion.
- *Intermediate 3D representation to decouple human and robot motion in VR teleoperation.* Instead of directly mirroring human head movements and streaming live robot camera views, we use an intermediate 3D scene representation. This representation enables real-time rendering of novel views based on the human's latest head pose, without requiring new observations from the robot. Consequently, the robot can be slowed down to reflect *aggregated* head movements rather than every motion. This asynchronous streaming, control, and rendering bypasses the need for low-latency robot actuation and data transmission.
- *Shared-observation teleoperation as a scalable way to capture active perception strategies.* Instead of hand-designing a gaze strategy, we let the policy learn the strategy directly from human demonstrations. By having the human use the same observation space as the robot—*seeing what the robot sees*—we effectively capture the human's complex perceptual strategies across task stages and scenarios. This enables the visuomotor policy to learn robust gaze behavior, even with straightforward behavior cloning.

To evaluate our proposed system, we perform experiments on three challenging, multi-stage bimanual manipulation tasks involving significant visual occlusions. These tasks include retrieving objects with interactive perception, rearranging cups in cluttered environments with active viewpoint switching, and precisely aligning objects using coordinated bimanual actions. Our experimental results highlight the critical role of active perception, with ViA outperforming baseline camera setups—such as wrist cameras and fixed chest cameras—by 45% in success rate. We also conducted a user study to validate the design of our teleoperation interface. Results are best viewed on our website: https://vision-in-action.github.io.

2 Related Work

Active Perception and Robot Necks. Active perception has a long-standing history in robotics and computer vision [1, 24, 25, 26, 18, 17, 19]. To investigate active perception, many artificial vision systems (*i.e.*, humanoid necks with varying numbers of degrees of freedom) have been developed [11, 12, 13, 27, 14, 28, 29, 30, 31]. In this paper, we use an off-the-shelf 6-DoF arm as

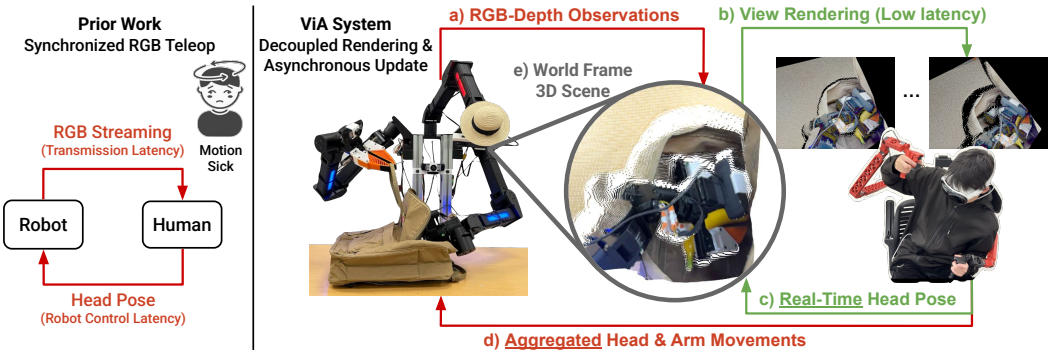

Figure 2: **VR Teleoperation Comparison.** [Left] Traditional RGB streaming suffers from motion-to-photon latency due to both RGB data transmission latency and robot control latency, often leading to VR motion sickness. [Right] Our system mitigates this by: (a, e) streaming a 3D point cloud in the world frame from RGB-D data, (b, c) performing real-time view rendering based on the user's latest head pose, and (d) asynchronously updating the robot's head and arm poses. This approach enables low-latency viewpoint updates for the user.

an active neck, with a camera mounted to the end effector. While the idea of using a robot arm as a neck has been explored in prior work [32, 33, 34, 15, 35], our approach integrates a novel teleoperation interface to directly control the robot neck. The majority of prior active vision systems use various heuristics (*e.g.*, hand-designed image filters or object detectors) to compute a measure of saliency for gaze guidance [20, 21, 22]. There are also works that formulate the next-best-view problem in terms of uncertainty reduction [36, 37, 23]. In these types of methods, the objective is defined purely around a perception problem, and do not consider manipulation. Recent works have also investigated active vision with reinforcement learning approaches [38, 39, 40, 41, 35, 42], though application of those methods to real-world systems remains challenging. In contrast, our work learns bimanual manipulation and active perceptual behavior directly from real-world human demonstrations, without any task-specific assumptions.

Teleoperation Systems. Recent teleoperation works [2, 3, 10, 43, 44, 45, 46] have highlighted the potential of scaling end-to-end visuomotor policy learning. However, existing approaches [2, 3, 47, 9] typically rely on wrist cameras [9, 48] or fixed third-person cameras [49], which fail to capture the active perceptual behaviors of humans. To overcome this human-robot observation mismatch, prior works [14, 15, 50, 51, 52, 53, 34, 54] have explored using VR to control an active head camera [14], providing immersive, first-person visual feedback through RGB video streaming. However, these direct camera teleoperation approaches [14, 52] often induce motion sickness [55], primarily due to motion-to-photon latency—the delay between the user's head movement and the corresponding visual update on the VR display [56]. To address this, our method introduces an intermediate 3D scene representation that enables real-time view rendering based on the user's latest head pose, significantly reducing motion-to-photon latency. Related to our approach, recent work [57] introduced a VR teleoperation system that uses radiance fields to render views from a reconstructed scene. However, unlike our approach, their system lacks physical camera control. Our approach, by contrast, allows users to purposefully control the camera in VR to maintain task-relevant visibility.

3 The Vision in Action System

The ViA system features a simple yet effective robotic neck design that allows the robot to mimic human whole-upper-body movements (§3.1). We introduce a 3D scene interface that renders views in real-time based on the user's latest head pose. This interface asynchronously updates the underlying 3D environment while allowing the user to purposefully control the robot's active camera (§3.2). Finally, we propose a visuomotor policy learning framework that leverages active perception (§3.3).

3.1 Hardware Design

Human active perception relies on coordinated movements of both the torso and neck to adjust head poses and acquire better viewpoints. However, the common approach of mounting a 2-DoF neck on a static torso [14] provides limited flexibility and is insufficient to replicate the full range of motion.

To address this, we use an off-the-shelf 6-DoF ARX5 robot arm as a robot neck. This high-DoF neck design allows the robot to mimic human-like head motions that naturally result from whole-upper-body movements. The active head camera streams real-time RGB, depth, and synchronized camera pose data. To meet these requirements, we use an iPhone 15 Pro [6], mounted on the end effector of the robot neck, as the system's primary visual sensor. To enable bimanual manipulation, we use two additional 6-DoF ARX5 robot arms [58] each equipped with a fin-ray parallel-jaw gripper. Each arm is mounted onto a custom 3D-printed shoulder structure.

3.2 Teleoperation Interface

To collect human demonstration data, we designed a teleoperation interface that simultaneously controls both robot arms and the active neck. For the arm teleoperation, we use a full-scale bimanual exoskeleton (inspired by GELLO [47]) that enables joint-to-joint mapping between the human user and the robot arms. For head teleoperation, we implemented a VR interface that allows the user to control the pose of the active head camera while observing visual feedback. Our choice of VR for the head interface was motivated by the need to precisely capture human perceptual strategies. By constraining the user to use the same observations as the robot, we can record visual attention patterns that contribute to successful task execution.

Challenge: *Exacerbated latency from physical movements.* In the VR literature, motion-to-photon latency, also known as end-to-end latency [56], refers to the delay between a user's head movement and the corresponding visual update on the display. High latency can cause discomfort or motion sickness. While today's consumer VR headsets achieve acceptable motion-to-photon latency (below 10 ms) for applications like games [59], robot teleoperation introduces an additional challenge—*robot control latency*. When users move their heads to teleoperate the robot's camera, there is a delay between the robot receiving and executing the command, causing the camera control to lag behind. This additional delay creates a mismatch between the user's head movement and visual feedback, leading to potential motion sickness.

Solution: *View decoupling through an intermediate 3D scene representation.* To overcome this challenge, we decouple the user's view from the robot's view using an intermediate 3D scene representation (Fig. 2). This allows the user's viewpoint to update **instantly** in response to head movements (via rendering), without waiting for the robot to physically match the requested viewpoint. While the rendered view may contain small regions with missing information (due to the delayed camera movement), *the rendered view stays aligned with the user's latest head pose*—a critical factor for preserving perceptual continuity and reducing discomfort. Concretely, the interface has three components:

- *Point cloud construction in the world frame.* We define the world frame W at the fixed base of the robot neck. Each RGB-D frame is transformed into this world frame using the camera intrinsics and the robot head pose (*i.e.*, camera extrinsics w.r.t. the world frame) at time t, denoted as ${}^{W}T_{\mathrm{H}}(t)$. This pose is computed by composing the iPhone's real-time relative pose with the initial robot head pose ${}^{W}T_{\mathrm{H}}(t_0)$, obtained from the robot neck's joint positions. The resulting point cloud ${}^{W}X(t)$ in the world frame serves as our intermediate 3D scene representation.
- *Low-latency view rendering.* From the point cloud ${}^{W}X(t)$, we render stereo RGB views for the VR display using the user's latest head pose in the world frame, denoted as ${}^{W}T_{\mathrm{user}}(t+k)$. This pose is computed by transforming the VR device's head pose into the world frame W with a height offset. This view rendering—where k denotes a short time interval—enables instant visual feedback for the user. Combined with a high refresh rate (roughly 150 Hz), our system ensures smooth viewpoint updates with minimal perceived latency.
- *Point cloud updating with aggregated head movements.* Finally, the robot head pose is updated to ${}^{W}T_{\mathrm{H}}(t+K)$ over a longer time interval K, using the aggregated user head pose, where K is determined by the robot's control latency and is much larger than the rendering interval k. Meanwhile, the point cloud is asynchronously updated with new RGB-D observations from the robot at a lower frequency (10 Hz).

Overall, this teleoperation interface balances low visual latency for the user (< 7 ms) with smooth action execution on the robot side (< 10 Hz control frequency), enabling effective and practical data collection for complex manipulation tasks.

3.3 Learning Active Perception for Bimanual Manipulation

We design a visuomotor policy network based on Diffusion Policy [10] that leverages our active head camera setup to learn from human demonstrations. The policy predicts bimanual arm actions for manipulation and neck actions that mimic human active perception behaviors, conditioned on visual and proprioceptive observations.

To enable coordinated head and arm movements, we represent the end-effector poses of the neck and arms in a common world frame. At each time step t, the policy receives the current RGB image observation $\mathbf{I}_t \in \mathbb{N}_0^{H\times W\times C}$ from the active head camera as the visual input, along with the proprioceptive state $\mathbf{P}_t \in \mathbb{R}^{23}$. This state includes the end-effector poses (position and quaternion) of the neck, left arm, and right arm ($\in \mathbb{R}^7$), as well as the two gripper widths (2 scalars).

We adopt a DINOv2 [60] pretrained ViT as the visual encoder for the RGB image $\mathbf{I}_t$ from the active head camera. The 384-dimensional classification token is extracted as a compact semantic representation of the visual scene. The policy outputs a sequence of future actions $\mathbf{A}_t = \{a_{t+1}, \ldots, a_{t+n_p}\} \in \mathbb{R}^{n_p\times 23}$, where each action consists of the future end-effector poses of the neck and arms in the world frame, as well as the gripper widths. Only the first $n_a \leq n_p$ actions are executed on the physical robot (via inverse kinematics). We use a prediction horizon of $n_p = 16$ and an execution horizon of $n_a = 8$, with the policy operating at 10 Hz.

4 Evaluation

We evaluated our system on three challenging multi-stage tasks (Fig. 3) to assess the effectiveness of various camera setups (§4.1), visual representations (§4.2), and teleoperation interface designs (§4.3). For each task, we report the stage-wise success, which is defined cumulatively (*i.e.*, success at each stage requires the successful completion of all preceding stages). Detailed task stage definitions can be found in the supplementary material.

Bag Task**: Object retrieval with interactive perception.* The robot must (1) open a bag, (2) peek inside to locate the target object, and (3) take it out. Success requires both ***active physical interaction (*i.e.*, opening the bag to reduce occlusion) and ***active head movement*** to inspect the bag's interior, demonstrating interactive perception. The wrist camera often suffers from limited visibility due to occlusions, whereas the active head camera can dynamically adjust its viewpoint to gather task-relevant information more effectively. We collected 150 demonstrations with five *training objects* (banana, carrot, dog, shoe, strawberry) and evaluated on two *unseen test objects* (a blue elephant, a green avocado) with 5 rollouts per object—10 trials in total. For both training and evaluation, a single object is placed in the bag per trial.

Cup Task**: Cup arrangement with active viewpoint switching.* As illustrated in Fig. 3, the robot must (1) find and pick up a cup from shelf A using its right hand, (2) hand it over to its left hand, and (3) place it on a saucer hidden beneath shelf B. Visual occlusion presents a significant challenge, requiring ***active viewpoint switching across different stages: the cup is positioned deep within shelf A, where upper tiers obstruct wrist cameras, while the saucer is positioned beneath shelf B. We collected 125 training demonstrations, with the cup randomly placed on either the upper or lower tier of shelf A and the saucer randomly positioned beneath shelf B. Demonstrations followed a consistent search strategy (lower tier first, then upper if needed). For evaluation, we used 10 test configurations, each run twice, resulting in 20 total rollouts.

Lime & Pot Task**: Bimanual coordination and precise alignment.* The robot must (1) find and place a lime into a pot, (2) lift the pot using both arms, and (3) precisely align it onto a trivet. Since the lime may appear on either side of the workspace, the robot must first ***coordinate and decide which arm to use for grasping. Lifting the pot requires ***bimanual grasping***, and the final ***precise alignment*** with the trivet is guided by the head camera to ensure precise placement. We collected

Figure 3: **Task Definitions.** We introduce three multi-stage tasks that highlight the critical role of active perception in everyday scenarios. [Left] Third-person view with red arrows indicating head movements and blue arrows indicating arm movements. [Middle] Active head camera views across task stages (upper row), and third-person view of robot actions (lower row). [Right] Test scenarios, including training and testing objects for the bag task, and different test configurations for the latter two tasks.

260 demonstrations for training. For evaluation, we fixed the pot position and tested 10 different lime and trivet configurations, each tested twice for 20 total rollouts.

4.1 Policy Learning Camera Setup Comparison

Camera Setups. We evaluate the effectiveness of active head camera setup by comparing it with two alternative camera configurations for policy learning (Fig. 4). During data collection, all camera streams are recorded. For training, we use different combinations of these views from the same set of demonstrations, enabling a fair comparison across camera setups. Visual representations are extracted using a DINOv2 pretrained ViT backbone [60].

- [**ViA (Ours)**]: Uses a single active head camera as the visual input. Details are described in §3.3.
- [**Active Head & Wrist Cameras**]: Combines the active head camera with two wrist cameras. Compared to [ViA], this setup includes additional wrist views as visual input. Although the teleoperator does not directly use these views, this comparison evaluates whether they provide additional useful information for policy learning.
- [**Chest & Wrist Cameras**]: Uses a fixed chest camera and two wrist cameras (omitting the neck). This is one of the most commonly used camera setups in current robotics systems [2, 3].

Results. As shown in Fig. 5, [ViA] consistently outperforms both alternative camera setups across all three tasks. Surprisingly, augmenting [ViA] with additional wrist camera observations ([Active Head & Wrist Cameras]) does not improve performance (a decrease of 18.33% on average). We

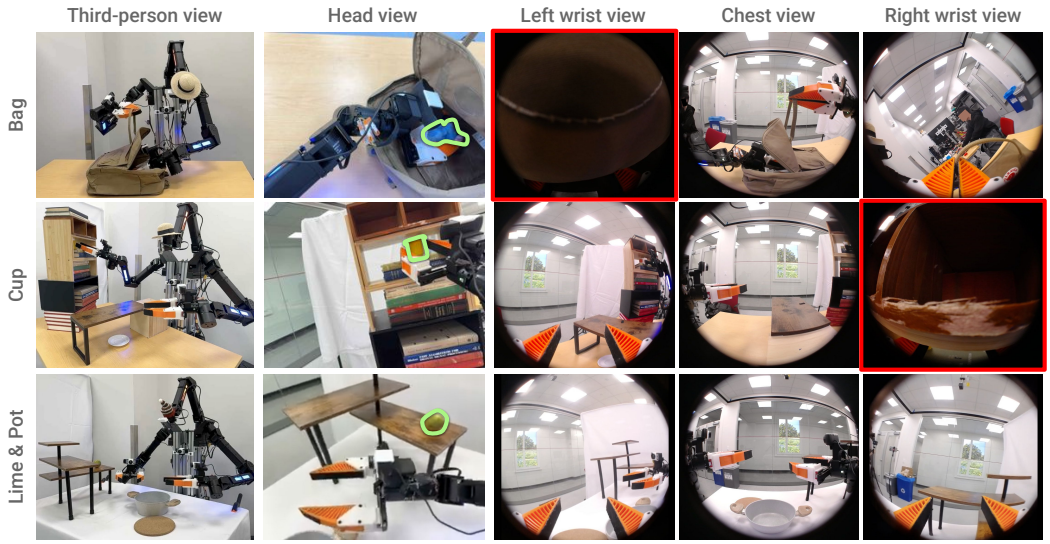

Figure 4: **Policy Learning Camera Setup Comparison.** [ViA] uses a single active head camera that dynamically adjusts its viewpoint to capture task-relevant visual information (*e.g.*, finding a cup hidden inside a shelf). In contrast, [Wrist & Chest cameras] policy often fails due to visual occlusions. For example, in the cup task, the right wrist camera's view is blocked by the upper shelf tier, resulting in insufficient visual cues for grasping. The chest camera also fails to capture task-relevant information due to its fixed viewpoint, even when equipped with a fisheye lens.

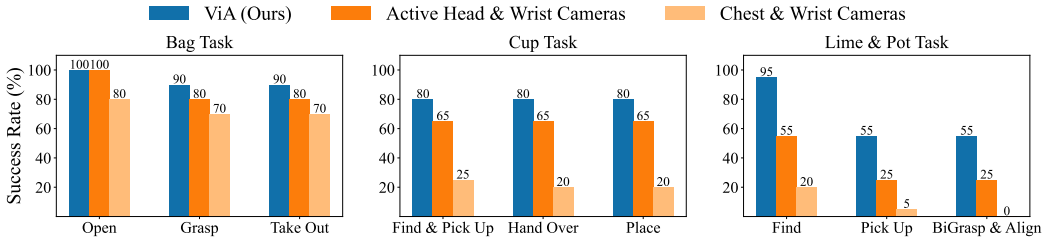

Figure 5: **Policy Learning Camera Setup Comparison Results.** We report stage-wise success rates across the three tasks to demonstrate the effectiveness of our active head camera [ViA] compared to two baseline configurations: [Active Head & Wrist Cameras] and [Chest & Wrist Cameras].

hypothesize several reasons for this outcome: First, the active head camera alone already provides sufficient information, as the teleoperator relies solely on this view to complete the task. Thus, the visual input from the head camera alone is already task-complete. Second, adding wrist cameras increases input dimensionality without necessarily contributing task-relevant information. Instead, the additional views may introduce redundant or noisy observations, especially due to frequent occlusions during manipulation. In a low-data regime like ours, the added complexity can hinder learning by increasing the risk of overfitting or distracting the model with less informative inputs.

Next, compared to the [Chest & Wrist Cameras] setup, it is clear that the chest and wrist cameras fail to provide sufficient task-relevant information. As shown in the second row of Fig. 4, the right wrist camera is completely occluded by the upper shelf tier during cup-grasping, while the fixed chest camera lacks visibility of the target objects altogether. In contrast, our active head camera dynamically adjusts its viewpoint, allowing the robot to gather more informative visual input and improve average task performance by 45%.

4.2 Policy Learning Visual Representation Comparison

Visual Representations. We compare [ViA] with two alternative visual representations for the policy. All policies use the same active head camera input as [ViA].

- [**ViA (Ours)**]: Uses a DINOv2 [60] vision backbone for image encoding. Details can be found in §3.3.

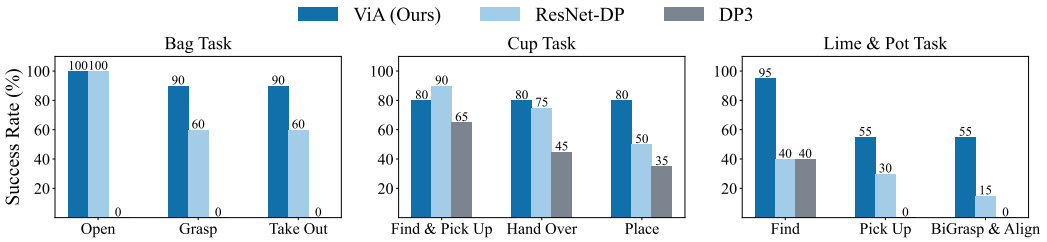

Figure 6: **Policy Learning Visual Representation Comparison Results.** We report stage-wise success rates across the three tasks to demonstrate the effectiveness of our method [ViA], in comparison to two baseline approaches: [ResNet-DP] and [DP3].

- [**ResNet-DP**]: A baseline using a ResNet-18 [61] backbone pretrained on ImageNet [62], integrated into diffusion policy. Input images are center-cropped to 1:1 aspect ratio and resized to 224×224, consistent with [ViA].
- [**DP3**] [63]: Uses world-frame point clouds (transformed from the active head camera) as visual input. The point cloud is cropped to the workspace and downsampled to 1,024 points. This model is trained from scratch.

Results. As shown in Fig. 6, our method—leveraging a pretrained DINOv2 ViT representation—achieves the highest final-stage success rate across all three tasks. Compared to the two baselines, [ViA] benefits from stronger semantic understanding enabled by the DINOv2 backbone. This allows the policy to actively *find* the object first before initiating arm actions. For example, in the lime & pot task, [ViA] is able to perform long-horizon active search to find the lime, before proceeding with manipulation. In contrast, a common failure mode of the [DP3] baseline is hallucination, where the policy misinterprets the scene and issues incorrect actions. For example, in the cup task, [DP3] often directs the arm to an empty section of the shelf, failing to identify the actual cup location. [DP3] also completely fails on the bag task due to the imprecise grasping of the bag handle in the open stage. We hypothesize that this is due to the limited semantic capacity of the [DP3] representation, which is trained from scratch and lacks pretrained visual priors.

4.3 Teleoperation Interface Comparison

In this experiment, we evaluate our VR teleoperation interface by comparing our system—which uses a point cloud rendering method—with a conventional RGB streaming approach [14, 15, 52]. We conducted a user study with 8 participants of varying levels of experience with VR and robot teleoperation. All participants were first-time users of both systems and were unaware of which system corresponded to our proposed design. For each participant, the order of system usage was randomized and labeled as System A and System B. Participants were asked to perform the cup task using both systems. Each session included a 5-minute practice period followed by a data collection phase in which participants provided 3 demonstrations. We recorded the completion time for each demonstration and gathered user feedback through a post-session experience survey.

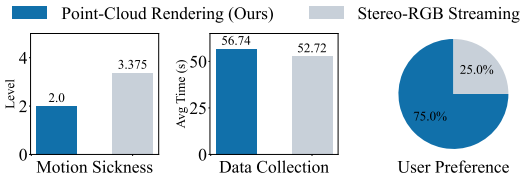

Figure 7: **Teleoperation Interface Comparison.** We evaluate our teleoperation interface design based on three metrics: reported levels of motion sickness, average duration to complete each demonstration, and overall user preference.

Results. As shown in Fig. 7, while our point cloud rendering method results in slightly longer data collection times compared to stereo RGB streaming, it significantly reduces motion sickness. As a result, 6 out of 8 participants reported a preference for our system.

5 Conclusion

The ViA system features a simple yet effective neck design that allows the robot to mimic human-like head movements. We developed a teleoperation interface that renders real-time views based on the user's latest head pose, while asynchronously updating the scene by controlling the robot's active head camera to gather task-relevant information. For evaluation, we introduced three challenging multi-stage tasks involving significant visual occlusion for policy learning. Experimental results highlight the importance of active perception, with ViA significantly outperforming baseline setups.

6 Limitations

This work explores how active perception can be learned from human demonstrations, taking an initial step toward that broader goal. While our system shows promising results, it also presents several limitations. Below, we outline three areas for future improvement and research.

Teleoperation Interface Design. Our 3D scene interface enables real-time view rendering from a point cloud, transformed from single-frame RGB-D data. However, due to noisy depth sensing and incomplete scene reconstruction, the resulting visualization can be lower in fidelity compared to traditional RGB video streaming. As a result, users often require practice to adapt and may find it challenging to perform fine-grained manipulation tasks. One possible future direction is to explore dynamic scene fusion and rendering techniques [64], which remain an important yet challenging problem. In addition, wearable devices such as AR glasses [4] hold great promise for capturing human active perception in everyday tasks, potentially removing the need for physical robot teleoperation during data collection.

Hardware Design. Using an off-the-shelf 6-DoF arm as a neck is a simple yet effective solution. However, this design may fall short in replicating the full complexity of human whole-body movements. We see exciting opportunities in optimizing hardware designs that enable more human-like behaviors and facilitate learning from humans [65, 66, 27]. We are also interested in upgrading our current tabletop systems to mobile manipulation platforms, where active perception becomes more challenging and better reflects real-world scenarios.

Policy Learning Design. There are several opportunities to improve the design choices in our current policy learning framework. First, we believe it is valuable to explore representation learning that fuses observations from all cameras into a shared space, rather than simply concatenating encoded features, which may improve overall task performance in the future. Second, our policy is not yet conditioned on language. Incorporating reasoning presents a promising direction, especially when combined with active perception. Natural language instructions often imply high-level goals, spatial cues, object relationships, or temporal dependencies. By conditioning policies on language, robots can better interpret human intent, dynamically adjust perception strategies, and disambiguate between similar visual scenes. Finally, tasks involving search—such as our Lime & Pot task—require memory. The robot must search for the lime before executing any arm actions, and ideally remember which areas have already been searched to avoid repetitive behavior. Our current policy learning framework does not support such memory capabilities.

Acknowledgments

This work was supported in part by the Toyota Research Institute, NSF Award #2143601, #2037101, and #2132519, the Sloan Foundation, Stanford Human-Centered AI Institute, and Intrinsic. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

We would like to thank ARX for the ARX robot hardware. We thank Yihuai Gao at Stanford for his help on the ARX robot arm controller SDK. We thank the help from Ge Yang at MIT and Xuxin Cheng at UCSD for their help and discussion of VR. We thank Max Du, Haochen Shi, Austin Patel, Zeyi Liu, Huy Ha, Mengda Xu, Yunfan Jiang, Ken Wang, Yanjie Ze for their helpful discussions. We thank all the volunteers who participated in and supported our user study.

References

[1] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. doi:10.1109/5.5968.

[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[3] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Conference on Robot Learning (CoRL)*, 2024.

[4] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. Egomimic: Scaling imitation learning via egocentric video, 2024. URL https://arxiv.org/abs/2410.24221.

[5] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[6] H. Etukuru, N. Naka, Z. Hu, S. Lee, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. M. Shafiullah. Robot utility models: General policies for zero-shot deployment in new environments. *arXiv preprint arXiv:2409.05865*, 2024.

[7] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[8] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024.

[9] K. Shaw, Y. Li, J. Yang, M. K. Srirama, R. Liu, H. Xiong, R. Mendonca, and D. Pathak. Bimanual dexterity for complex tasks. In *8th Annual Conference on Robot Learning*, 2024.

[10] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[11] S. Shigemi. *ASIMO and Humanoid Robot Research at Honda*, pages 55–90. Springer Netherlands, Dordrecht, 2019. ISBN 978-94-007-6046-2. doi:10.1007/978-94-007-6046-2_9. URL https://doi.org/10.1007/978-94-007-6046-2_9.

[12] K. Kaneko, K. Harada, F. Kanehiro, G. Miyamori, and K. Akachi. Humanoid robot hrp-3. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2471–2478, 2008. doi:10.1109/IROS.2008.4650604.

[13] M. Elmogy, C. Habel, and J. Zhang. Online motion planning for hoap-2 humanoid robot navigation. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3531–3536, 2009. doi:10.1109/IROS.2009.5354572.

[14] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[15] B. Sen, M. Wang, N. Thakur, A. Agarwal, and P. Agrawal. Learning to look around: Enhancing teleoperation and learning with a human-like actuated neck, 2024. URL https://arxiv.org/abs/2411.00704.

[16] W. Zheng, Y. Sun, H. Wu, H. Sun, and D. Zhang. The interaction of top–down and bottom–up attention in visual working memory. *Scientific Reports*, 14(1):17397, 2024.

[17] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[18] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[19] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1–2):507 – 545, 1995. ISSN 0004-3702. doi:http://dx.doi.org/10.1016/0004-3702(95)00025-9. Special Volume on Computer Vision.

[20] B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic. An active vision system for detecting, fixating and manipulating objects in the real world. *The International Journal of Robotics Research*, 29(2-3):133–154, 2010. doi:10.1177/0278364909346069.

[21] M. Grotz, T. Habra, R. Ronsse, and T. Asfour. Autonomous view selection and gaze stabilization for humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1427–1434, 2017. doi:10.1109/IROS.2017.8205944.

[22] J. Bohg, K. Welke, B. León, M. Do, D. Song, W. Wohlkinger, M. Madry, A. Aldóma, M. Przybylski, T. Asfour, H. Martí, D. Kragic, A. Morales, and M. Vincze. Task-based grasp adaptation on a humanoid robot. *IFAC Proceedings Volumes*, 45(22): 779–786, 2012. ISSN 1474-6670. doi:https://doi.org/10.3182/20120905-3-HR-2030.00174. URL https://www.sciencedirect.com/science/article/pii/S1474667016337041. 10th IFAC Symposium on Robot Control.

[23] R. Pito. A solution to the next best view problem for automated surface acquisition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1016–1030, 1999. doi:10.1109/34.799908.

[24] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos. Revisiting active perception. *CoRR*, abs/1603.02729, 2016.

[25] D. H. Ballard. Animate vision. *Artificial intelligence*, 48(1):57–86, 1991.

[26] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1:333 – 356, 1988.

[27] H. Shi, W. Wang, S. Song, and C. K. Liu. Toddlerbot: Open-source ml-compatible humanoid platform for loco-manipulation, 2025. URL https://arxiv.org/abs/2502.00893.

[28] L. Righetti, M. Kalakrishnan, P. Pastor, J. Binney, J. Kelly, R. Voorhies, G. Sukhatme, and S. Schaal. An autonomous manipulation system based on force control and optimization. *Autonomous Robots*, 36:11–30, 01 2014. doi:10.1007/s10514-013-9365-9.

[29] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann. The karlsruhe humanoid head. In *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, pages 447–453, 2008. doi:10.1109/ICHR.2008.4755993.

[30] K. Pahlavan and J.-O. Eklundh. A head-eye system—analysis and design. *CVGIP: Image Understanding*, 56(1):41–56, 1992. ISSN 1049-9660. doi:https://doi.org/10.1016/1049-9660(92)90084-G. URL https://www.sciencedirect.com/science/article/pii/104996609290084G. Purposive, Qualitative, Active Vision.

[31] D. Kappler, F. Meier, J. Issac, J. Mainprice, C. G. Cifuentes, M. Wüthrich, V. Berenz, S. Schaal, N. Ratliff, and J. Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018. doi:10.1109/LRA.2018.2795645.

[32] T. Olson and R. Potter. Real time vergence control. In *Proceedings CVPR '89: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 404–409, 1989. doi:10.1109/CVPR.1989.37878.

[33] J. L. Crowley, P. Bobet, and M. Mesrabi. Gaze control for a binocular camera head. In G. Sandini, editor, *Computer Vision — ECCV'92*, pages 588–596, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg. ISBN 978-3-540-47069-4.

[34] C. Lenz and S. Behnke. Bimanual telemanipulation with force and haptic feedback through an anthropomorphic avatar system. *Robotics and Autonomous Systems*, 161:104338, 2023. ISSN 0921-8890. doi:https://doi.org/10.1016/j.robot.2022.104338. URL https://www.sciencedirect.com/science/article/pii/S0921889022002275.

[35] J. Lv, Y. Feng, C. Zhang, S. Zhao, L. Shao, and C. Lu. SAM-RL: Sensing-Aware Model-Based Reinforcement Learning via Differentiable Physics-Based Simulation and Rendering. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi:10.15607/RSS.2023.XIX.040.

[36] R. Zeng, Y. Wen, W. Zhao, and Y.-J. Liu. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 6(3):225–245, Sept. 2020. ISSN 2096-0662. doi:10.1007/s41095-020-0179-3. URL https://doi.org/10.1007/s41095-020-0179-3.

[37] M. Krainin, B. Curless, and D. Fox. Autonomous generation of complete 3d object models using next best view manipulation planning. In *2011 IEEE International Conference on Robotics and Automation*, pages 5031–5037, 2011. doi:10.1109/ICRA.2011.5980429.

[38] S. Dass, J. Hu, B. Abbatematteo, P. Stone, and R. Martín-Martín. Learning to look: Seeking information for decision making via policy factorization. *arXiv preprint arXiv:2410.18964*, 2024.

[39] R. Cheng, A. Agarwal, and K. Fragkiadaki. Reinforcement learning of active vision for manipulating objects under occlusions, 2019. URL https://arxiv.org/abs/1811.08067.

[40] J. Shang and M. S. Ryoo. Active vision reinforcement learning under limited visual observability, 2023. URL https://arxiv.org/abs/2306.00975.

[41] D. Jayaraman and K. Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks, 2017. URL https://arxiv.org/abs/1709.00507.

[42] S. Uppal, A. Agarwal, H. Xiong, K. Shaw, and D. Pathak. Spin: Simultaneous perception interaction and navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18133–18142, June 2024.

[43] J. Wu, W. Chong, R. Holmberg, A. Prasad, Y. Gao, O. Khatib, S. Song, S. Rusinkiewicz, and J. Bohg. Tidybot++: An open-source holonomic mobile manipulator for robot learning. In *Conference on Robot Learning*, 2024.

[44] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

[45] Y. Jiang, R. Zhang, J. Wong, C. Wang, Y. Ze, H. Yin, C. Gokmen, S. Song, J. Wu, and L. Fei-Fei. Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities, 2025. URL https://arxiv.org/abs/2503.05652.

[46] X. Xu, D. Bauer, and S. Song. Robopanoptes: The all-seeing robot with whole-body dexterity. *arXiv preprint arXiv:2501.05420*, 2025.

[47] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.

[48] Y. Liu, X. Xu, W. Chen, H. Yuan, H. Wang, J. Xu, R. Chen, and L. Yi. Enhancing generalizable 6d pose tracking of an in-hand object with tactile sensing. *IEEE Robotics and Automation Letters*, 9(2):1106–1113, 2023.

[49] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024.

[50] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2023.

[51] Y. Liu, S. Mu, X. Chao, Z. Li, Y. Mu, T. Chen, S. Li, C. Lyu, X.-p. Zhang, and W. Ding. Avr: Active vision-driven robotic precision manipulation with viewpoint and focal length optimization. *arXiv preprint arXiv:2503.01439*, 2025.

[52] I. Chuang, A. Lee, D. Gao, M.-M. Naddaf-Sh, and I. Soltani. Active vision might be all you need: Exploring active vision in bimanual robotic manipulation, 2025. URL https://arxiv.org/abs/2409.17435.

[53] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635, 2018. doi:10.1109/ICRA.2018.8461249.

[54] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu. Generalizable humanoid manipulation with 3d diffusion policies, 2025. URL https://arxiv.org/abs/2410.10803.

[55] U. A. Chattha, U. I. Janjua, F. Anwar, T. M. Madni, M. F. Cheema, and S. I. Janjua. Motion sickness in virtual reality: An empirical evaluation. *IEEE Access*, 8:130486–130499, 2020. doi:10.1109/ACCESS.2020.3007076.

[56] J. Zhao, R. S. Allison, M. Vinnikov, and S. Jennings. Estimating the motion-to-photon latency in head mounted displays. In *2017 IEEE Virtual Reality (VR)*, pages 313–314, 2017. doi:10.1109/VR.2017.7892302.

[57] M. Wilder-Smith, V. Patil, and M. Hutter. Radiance fields for robotic teleoperation. *arXiv*, 2024.

[58] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song. UMI on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. In *Proceedings of the 2024 Conference on Robot Learning*, 2024.

[59] M. Yang, J. Zhang, and L. Yu. Perceptual tolerance to motion-to-photon latency with head movement in virtual reality. In *2019 Picture Coding Symposium (PCS)*, pages 1–5, 2019. doi:10.1109/PCS48520.2019.8954518.

[60] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[61] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[63] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy. *arXiv e-prints*, pages arXiv–2403, 2024.

[64] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, June 2024.

[65] X. Xu, H. Ha, and S. Song. Dynamics-guided diffusion model for robot manipulator design. *arXiv preprint arXiv:2402.15038*, 2024.

[66] R. Schneider, D. Honerkamp, T. Welschehold, and A. Valada. Task-driven co-design of mobile manipulators. *arXiv preprint arXiv:2412.16635*, 2024.